

NEURAL NETWORK INSTRUCTION SET ARCHITECTURE

BACKGROUND

[0001] This specification relates to an instruction set for computation of Deep Neural Network (“DNN”).

[0002] Neural networks are machine learning models that employ one or more layers of models to generate an output, e.g., a classification, for a received input. Some neural networks include one or more hidden layers in addition to an output layer. The output of each hidden layer is used as input to the next layer in the network, i.e., the next hidden layer or the output layer of the network. Each layer of the network generates an output from a received input in accordance with current values of a respective set of parameters.

[0003] Some neural networks include one or more convolutional neural network layers. Each convolutional neural network layer has an associated set of kernels. Each kernel includes values established by a neural network model created by a user. In some implementations, kernels identify particular image contours, shapes, or colors. Kernels can be represented as a matrix structure of weight inputs. Each convolutional layer can also process a set of activation inputs. The set of activation inputs can also be represented as a matrix structure.

SUMMARY

[0004] One innovative aspect of the subject matter described in this specification can be embodied in a computer-implemented method for accelerating tensor computations for a neural network having multiple neural network layers. The method includes receiving, by a processing unit of a hardware computing system, a single instruction that specifies parameters for performing a tensor computation for a neural network layer of the multiple neural network layers, wherein the data values comprise at least one value that specifies a type of the neural network layer. In response to receiving the single instruction, executing a loop nest comprising a plurality of loops, wherein a structure of the loop nest is defined at least in part by the type of the neural network layer. The method may include, performing, by the processing unit, the tensor computation to generate an output for the neural network layer in response to executing the loop nest. These and other implementations can each optionally include one or more of the following features. For example, the tensor computation can be at least a portion of a computation of a neural network layer. The parameters specified by the instruction may comprise a parameter that specifies a type of the neural network layer, and the structure of the loop nest can be defined at least in part by the type of the neural network

layer. Thereby, executing a loop nest comprising a plurality of loops may refer to performing a traversal over elements of a tensor in an order specified by a structure of nested loops, wherein said structure is specified e.g. by a depth of the loop nesting, and start and end indices, the stride and the direction of each loop respectively.

executing a loop nest comprising a plurality of loops, wherein a structure of the loop nest is defined based on one or more of the data values of the instruction.

[0009] Yet another innovative aspect of the subject matter described in this specification can be embodied in a non-transitory computer-readable storage medium comprising instructions executable by a processing unit which, upon such execution, causes the processing unit to perform the aforementioned computer-implemented method.

[0010] Other implementations of this and other aspects include corresponding systems, apparatus, and computer programs, configured to perform the actions of the methods, encoded on computer storage devices. A system of one or more computers can be so configured by virtue of software, firmware, hardware, or a combination of them installed on the system that in operation cause the system to perform the actions. One or more computer programs can be so configured by virtue of having instructions that, when executed by data processing apparatus, cause the apparatus to perform the actions.

[0011] The subject matter described in this specification can be implemented in particular embodiments so as to realize one or more of the following advantages. A processing unit can initiate programs to iterate deeply-nested loops with one instruction. The processing unit can execute received instructions encoded with data values for performing a tensor computation. Computation bandwidth of the processing unit is increased by reducing the number of instructions that the processor is required to execute when traversing a tensor to perform one or more computations. Instructions for performing tensor computations for a given neural network layer can be encoded and distributed amongst one or more computing systems of an example hardware computing system. Distribution of the encoded instructions to the various compute systems allows for increased computation bandwidth within a single system. Instruction quantity in a compute system is reduced because a single system is responsible only for a subset of the total computations needed for a given tensor.

[0012] The details of one or more implementations of the subject matter described in this specification are set forth in the accompanying drawings and the description below. Other

CLAIMS

1. A computer-implemented method for accelerating tensor computations for a neural network having multiple neural network layers, the method comprising:
 - receiving, by a processing unit of a hardware computing system, a single instruction that specifies data values for performing a tensor computation for a neural network layer of the multiple neural network layers, wherein the data values comprise at least one value that specifies a type of the neural network layer;
 - in response to receiving the single instruction, executing a loop nest comprising a plurality of loops, wherein a structure of the loop nest is defined at least in part by the type of the neural network layer; and
 - performing, by the processing unit, the tensor computation to generate an output for the neural network layer in response to executing the loop nest.
2. The method of claim 1, wherein the single instruction causes the processing unit to access at least one element of a dimension of a tensor, the element being a part of at least one index used in executing the loop nest during performance of the tensor computation.
3. The method of claim 1, wherein the single instruction causes the processing unit to access at least one memory address of an array in a storage medium, the memory address of the array comprising a variable that is read by the processing unit during performance of the tensor computation.
4. The method of claim 3, wherein performing the tensor computation comprises, providing, by the processing unit, at least one control signal to a tensor traversal unit (TTU) to cause the TTU to emit loop indices used in executing the loop nest during performance of the tensor computation.
5. The method of claim 3, further comprising, providing, by the processing unit, at least one control signal to a tensor traversal unit (TTU) to cause an array reference of the TTU to generate an address for a referenced array element used in executing the loop nest during performance of the tensor computation.

6. The method of claim 5, wherein the single instruction indicates a first TTU counter that is summed with a second TTU counter to generate an address for an array reference associated with the TTU.
7. The method of claim 1, wherein performing the tensor computation comprises, executing, by the processing unit, a first sync procedure that manages one or more operands associated with performance of the tensor computation, wherein managing an operand comprises stalling one or more loop nests based on a sync flag condition.
8. The method of claim 7, wherein performing the tensor computation comprises, executing, by the processing unit, a second sync procedure that manages incrementing a counter associated with a characteristic of the loop nest.
9. An electronic system, comprising:
a processing unit of a hardware computing system, the processing unit disposed in in the electronic system, the processing unit including one or more processing devices; and
one or more machine-readable storage devices for storing instructions that are executable by the one or more processing devices to perform operations for accelerating tensor computations for a neural network having multiple neural network layers, the operations comprising:
receiving, by the processing unit, a single instruction that specifies data values for performing a tensor computation for a neural network layer of the multiple neural network layers, wherein the data values comprise at least one value that specifies a type of the neural network layer;
in response to receiving the single instruction, executing a loop nest comprising a plurality of loops, wherein a structure of the loop nest is defined at least in part by the type of the neural network layer; and
performing, by the processing unit, the tensor computation to generate an output for the neural network layer in response to executing the loop nest.
10. The electronic system of claim 9, wherein the single instruction causes the processing unit to access at least one element of a dimension of a tensor, the element being a part of at least one index used in executing the loop nest during performance of the tensor computation.

11. The electronic system of claim 9, wherein the single instruction causes the processing unit to access at least one memory address of an array in a storage medium, the memory address of the array comprising a variable that is read by the processing unit during performance of the tensor computation.
12. The electronic system of claim 11, wherein performing the tensor computation comprises, providing, by the processing unit, at least one control signal to a tensor traversal unit (TTU) to cause the TTU to emit loop indices used in executing the loop nest during performance of the tensor computation.
13. A non-transitory computer-readable storage medium comprising instructions executable by a processing unit which, upon such execution, causes the processing unit to perform the method of any of claims 1 to 8.